

Evaluation of Multiple Choice Questions by Applying Item Response Theory for OUM Test Items.

MCQ



- Mei 2016 - 129 subjek MCQ
- September 2016 - 157 subjek MCQ
- January 2017 - 182 subjek MCQ

Item Analysis



Item analysis is the process of collecting, summarising and using information from students' responses to make decisions about each item.

Nitko, 2004

2 Approaches



- **Classical test theory (CTT)**
- **Modern test theory;
(Item Response Theory (IRT),
the Rasch model)**

CTT



- Student's ability is based on raw score
- Examinee score are test dependent.
Examinee may obtain higher scores on an easier test and lower score on a harder test

Test statistic & item



- **Item difficulty : (p-value, difficulty measure)**

Proportion of candidates getting an item correct

- **Item discrimination: (D)**

How each item discriminate between high performers and low performers

Formula for p-value



P-value = number of students choosing the correct answer
total number of students answering the item

$$IK = \frac{N^*}{N}$$

N^* : *bilangan calon yang memilih kunci (jawapan betul)*

N : *jumlah calon yang cuba menjawab item itu*

The popular index to show the difficulty level of dichotomous item is p (*Proportion Correct Index*)

For example if an item is administered to 100 candidates, and 70 got it right, in this context the p -value is 0.7.

Another example if an item is administered to 121 candidates and 68 got it right, the p -value is 0.56
($68/121=0.56$)

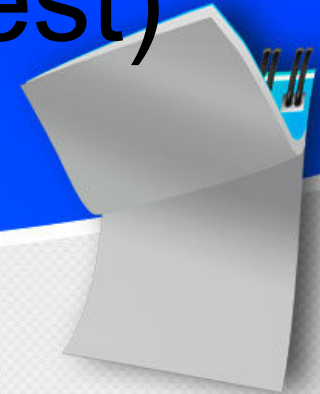
Item Analysis

Subject Code : BBMP1103	Subject: BBMP1103	Test Date: 2/ 9/2016								
Sorted by : ID Objective Mark : 40% Passing Mark : 40%	No. of Groups (Classes): 1	No. of Students: 609								
No. Topic	Omit	Multi	Correct	A	B	C	D	E	T	F

PI = Passing Index
DI = Discrim. Index

Q 1 Topic #01	1	0	478								PI = 0.79 DI = +0.29
Q 2 Topic #01	0	1	335								PI = 0.58 DI = +0.36
Q 3 Topic #01	0	0	289								PI = 0.48 DI = +0.25
Q 4 Topic #01	1	0	242								PI = 0.40 DI = +0.31
Q 5 Topic #01	0	0	378								PI = 0.63 DI = +0.58
Q 6 Topic #01	1	0	431								PI = 0.71 DI = +0.48
Q 7 Topic #01	0	0	423								PI = 0.70 DI = +0.60
Q 8 Topic #01	1	0	295								PI = 0.49 DI = +0.68
Q 9 Topic #01	0	0	310								PI = 0.51 DI = +0.76
Q10 Topic #01	1	0	330								PI = 0.55

Purposes of Designing (Test) Instruments



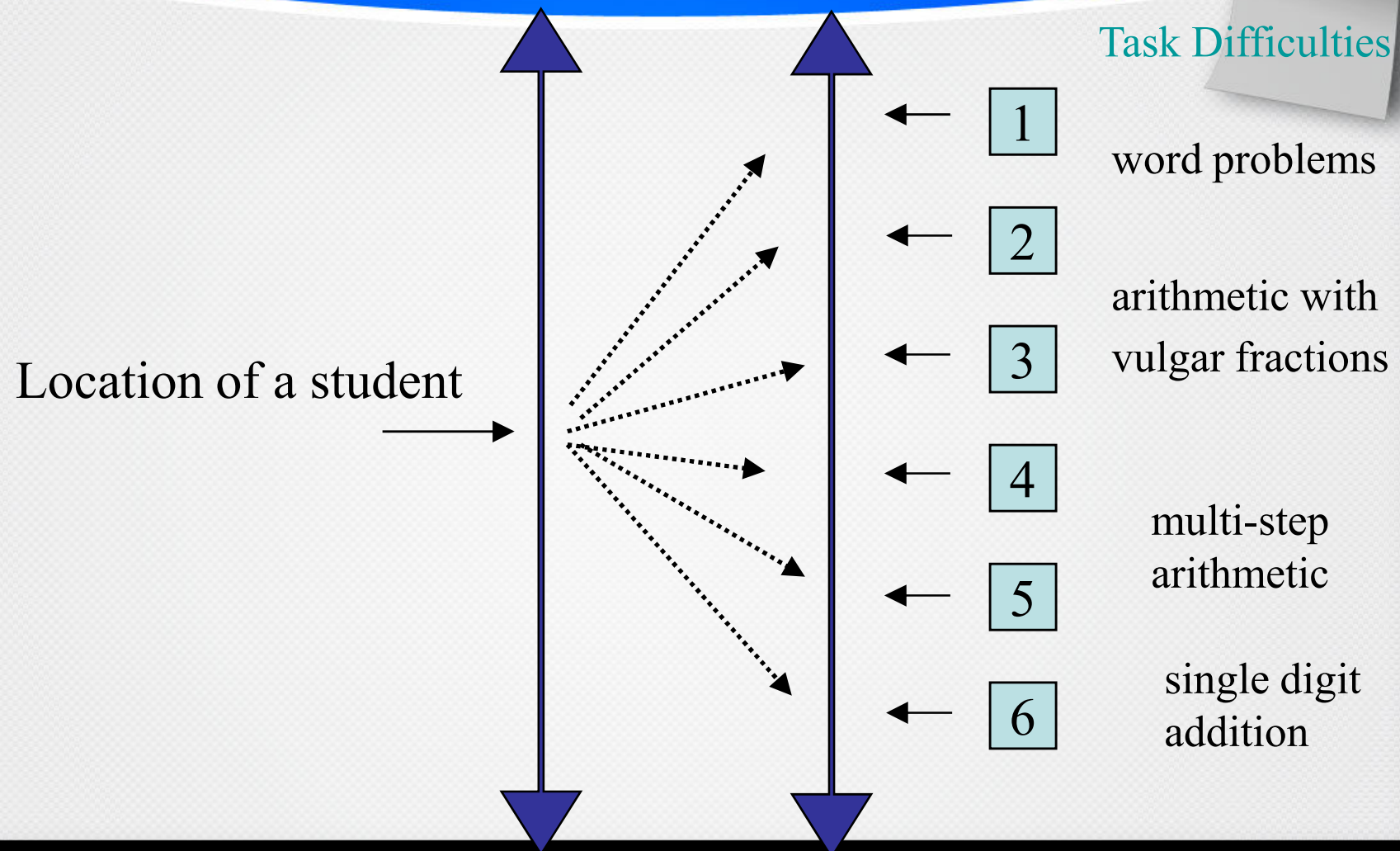
- To find out about something.
- What is it that we want to find out? About the items, or about the people?
- In general, “measurement” is concerned with finding out about characteristics (latent traits) of people. The items are instrumental (or incidental) to achieve the measurement.
- Measurement as measuring latent traits of people.

IRT

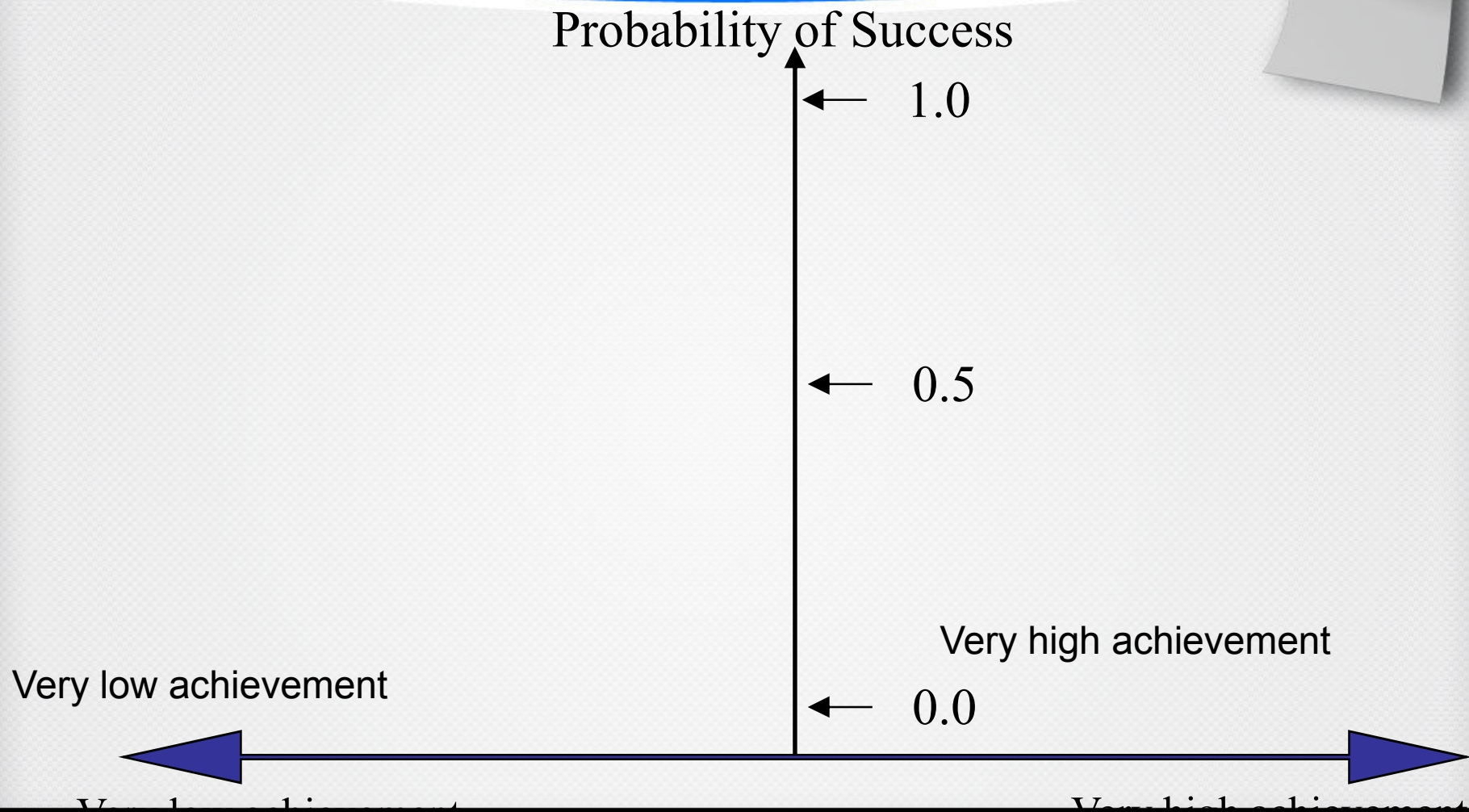


- IRT models give the *probability* of success of a person on an items.
- IRT models are not deterministic, but probabilistic.
- Given the item difficulty and person ability, one can compute the probability of success for each person on each item.

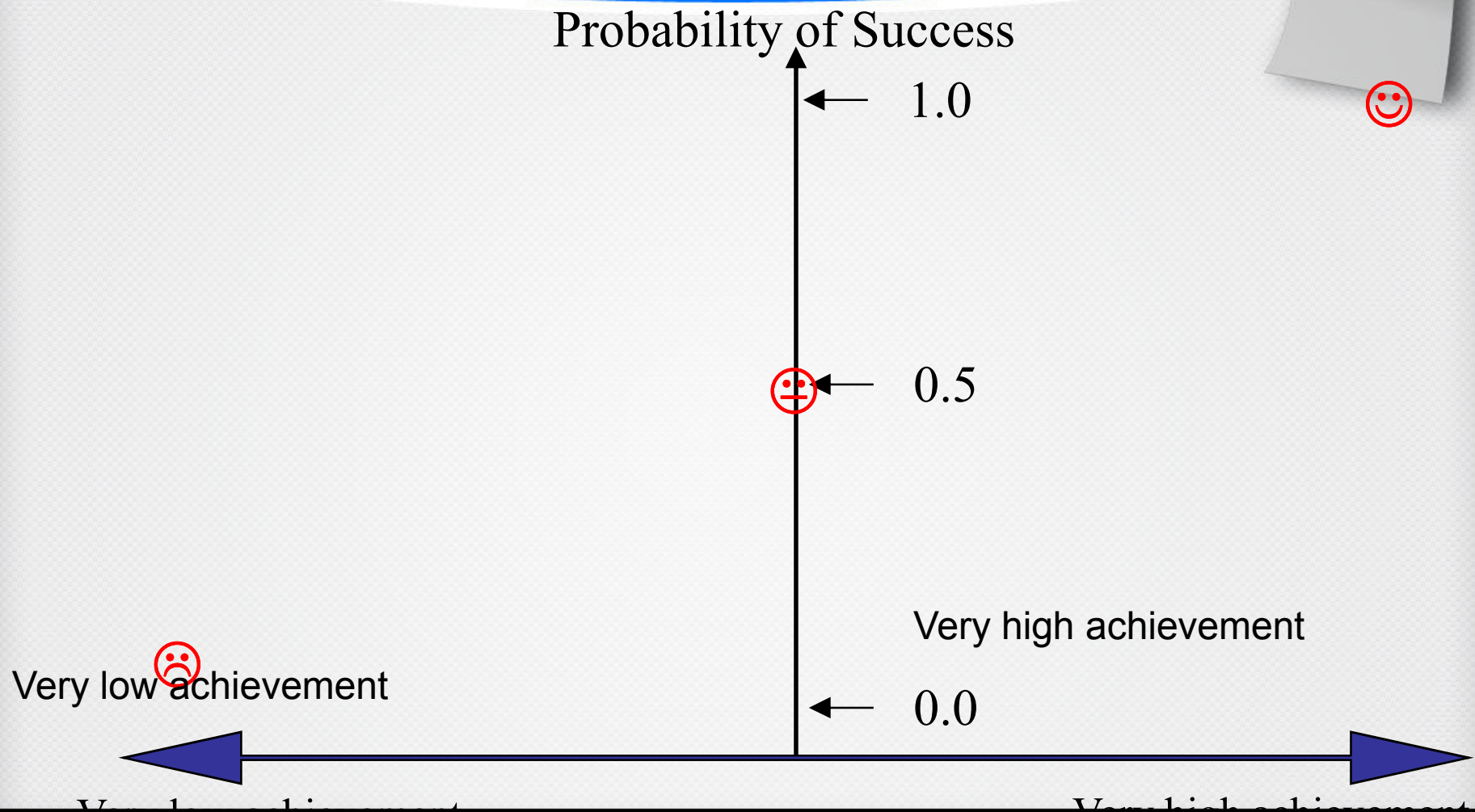
Comparing Students and Items



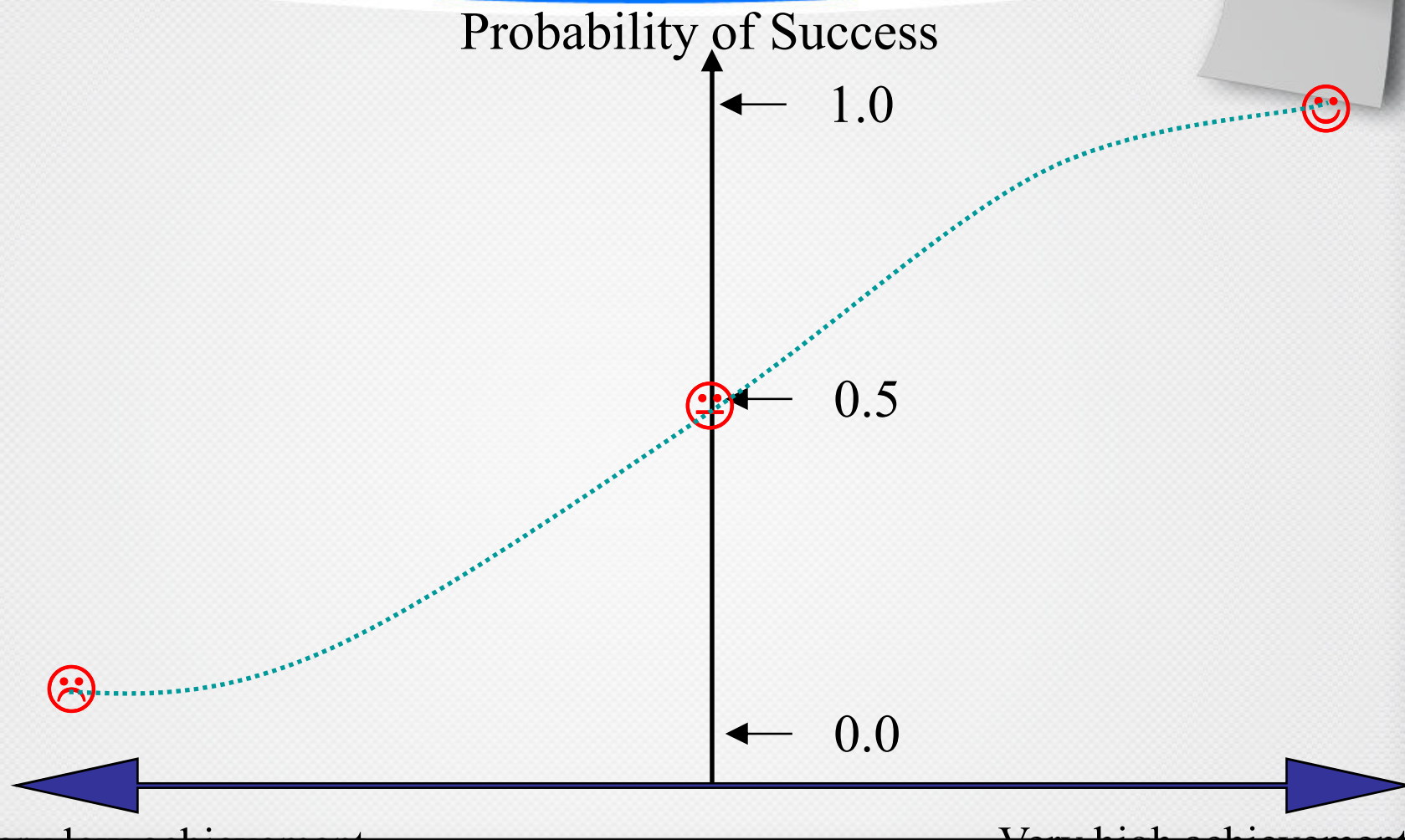
Building a Model



Imagine a middle difficulty task



Item Characteristic Curve



IRT



- 1 Parameter (Rasch)-difficulty
- 2 parameter- difficulty and guessing
- 3 parameter-discrimination, difficulty and guessing

Statistics



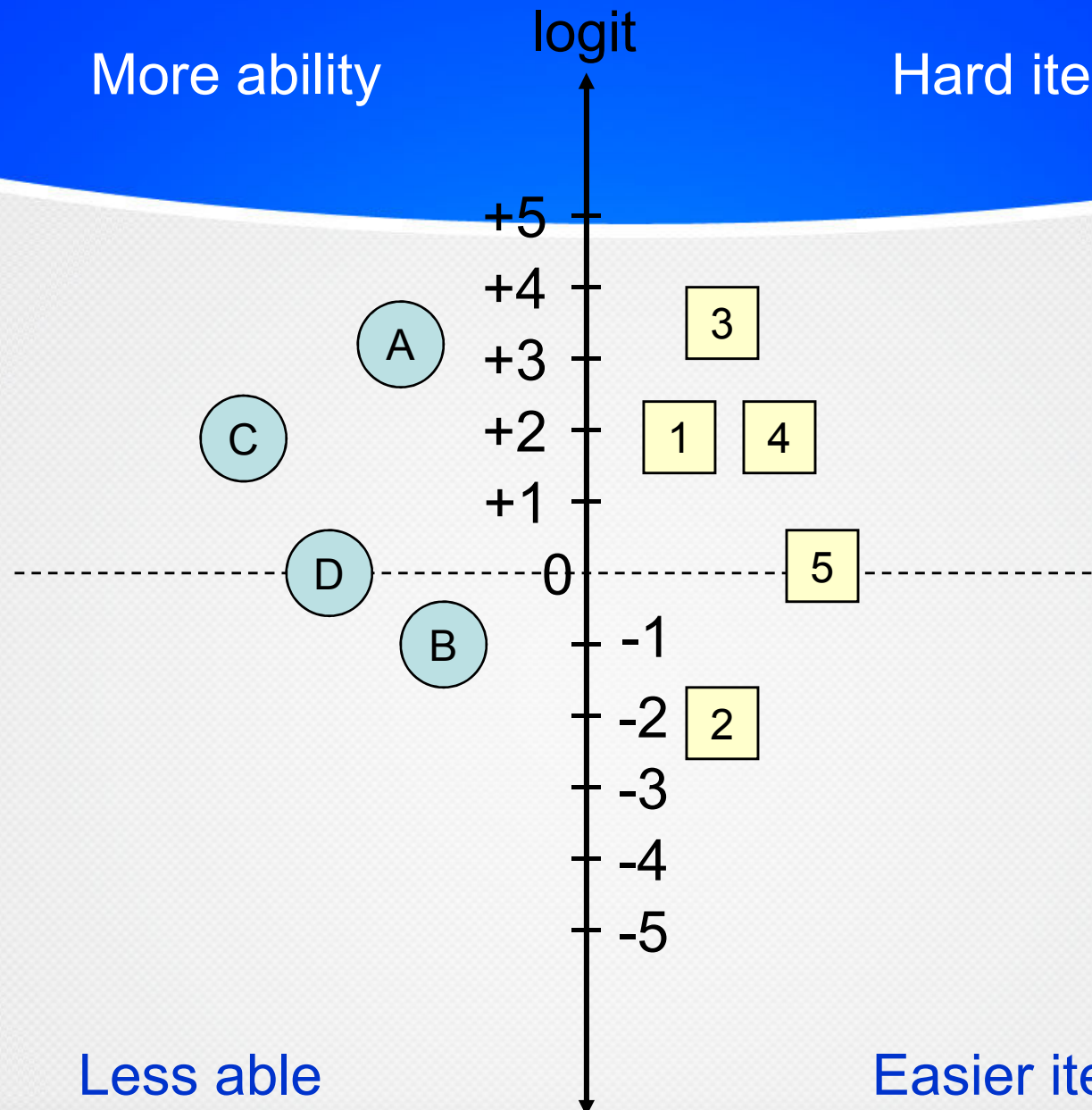
- Measurement for item (logit)
- Measurement for ability (logit)

Person location

Item location

More ability

Hard items



Less able

Easier items

Mathematics for Management Mei Semester 2016



- . 609 students**
- 40 multiple choice questions**
- WINSTEPS version 3.57**

Difficulty
Index

Discrimination
index

Reliability index

*Polarity

*Fit
analysis

Factor analysis/

*PCA / CFA

*Targetting

Item analysis

```
graph TD; IA([Item analysis]) --> DI[Difficulty Index]; IA --> DisI[Discrimination index]; IA --> RI[Reliability index]; IA --> P[*Polarity]; IA --> FA[*Fit analysis]; IA --> FAF[Factor analysis/ *PCA / CFA]; IA --> T[*Targetting];
```


Item properties



1. **Reliability-** the ordering of item difficulty is replicable with other comparable examinee (0.9-0.7)
1. **Separation Index for item** - indicates the number of difficulty levels of the test (3)
2. **Separation index for student-** number of groups of student can be separated according to ability
3. **Polarity(PMC)** – should have positive value, they are measuring the same construct(1- 0.36)
4. **Fit analysis-**discrepancies between the data and model expectation (0.7-1.3)
5. **Unidimensionality** – PCA not greater than 3 units

Reliability and Item separation

SUMMARY OF 40 MEASURED (NON-EXTREME) items

	RAW			MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR		MNSQ	ZSTD	MNSQ	ZSTD
MEAN	348.9	607.0	.00	.10		1.00	-.2	.99	-.2
S.D.	107.8	.0	.95	.01		.14	3.8	.21	3.4
MAX.	538.0	607.0	2.09	.13		1.39	8.7	1.53	8.6
MIN.	119.0	607.0	-1.90	.09		.77	-8.4	.64	-6.6
REAL RMSE	.10	ADJ.SD	.94	SEPARATION	9.51 item	RELIABILITY	.99		
MODEL RMSE	.10	ADJ.SD	.94	SEPARATION	9.75 item	RELIABILITY	.99		
S.E. OF item MEAN = .15									

UMEAN=.000 USCALE=1.000

item RAW SCORE-TO-MEASURE CORRELATION = -1.00

TABLE 3.1 BBMP1103

ZOU778ws.txt Feb 23 16:49 2017

INPUT: 609 students, 40 items MEASURED: 609 students, 40 items, 2 CATS 3.57.3

SUMMARY OF 607 MEASURED (NON-EXTREME) students

	RAW		MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	23.0	40.0	.44	.38	1.00	.0	.99	.0
S.D.	6.6	.0	.96	.08	.16	1.1	.27	1.1
MAX.	39.0	40.0	4.07	1.02	1.62	3.9	2.08	4.2
MIN.	9.0	40.0	-1.46	.35	.64	-3.3	.36	-2.9
REAL RMSE	.40	ADJ.SD	.88	SEPARATION	2.19	student	RELIABILITY	.83
MODEL RMSE	.39	ADJ.SD	.88	SEPARATION	2.26	student	RELIABILITY	.84
S.E. OF student MEAN = .04								

MAXIMUM EXTREME SCORE: 2 students

Unidimensionality – PCA not greater than 3 units

Table 3 - DIMENSIONALITY CHECK FOR BBMP1103

Dec 8 17:51

2016

INPUT: 609 students, 40 items MEASURED: 609 students, 40 items, 2 CATS 3.57.3

PRINCIPAL COMPONENTS (STANDARDIZED RESIDUAL) FACTOR PLOT

Factor 1 extracts 3.1 units out of 40 units of item residual variance noise.

Yardstick (variance explained by measures)-to-This Factor ratio: 10.3:1

Yardstick-to-Total Noise ratio (total variance of residuals): .8:1

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		Empirical	Modeled
Total variance in observations	=	72.1 100.0%	100.0%
Variance explained by measures	=	32.1 44.5%	44.2%
Unexplained variance (total)	=	40.0 55.5%	55.8%
Unexpl var explained by 1st factor	=	3.1 4.3%	

Table 2 - POLARITY CHECK FOR BBMP1103

Dec 8 17:51 2016

INPUT: 609 students, 40 items MEASURED: 609 students, 40 items, 2 CATS 3.57.3

student: REAL SEP.: 2.19 REL.: .83 ... item: REAL SEP.: 9.51 REL.: .99

item STATISTICS: CORRELATION ORDER

ENTRY	RAW			MODEL	INFIT		OUTFIT		PTMEA	
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	item
20	218	607	1.09	.09	1.39	8.7	1.53	8.6	.05	20
35	192	607	1.32	.10	1.27	5.5	1.39	5.8	.16	35
40	316	607	.30	.09	1.22	7.1	1.31	6.0	.16	40
18	213	607	1.13	.09	1.24	5.5	1.35	5.8	.18	18
16	530	607	-1.77	.13	1.02	.2	1.14	.9	.19	16
15	387	607	-.26	.09	1.13	3.9	1.24	3.5	.21	15
36	298	607	.44	.09	1.17	5.2	1.20	4.2	.23	36
19	469	607	-1.01	.10	1.03	.7	1.20	1.9	.24	19
37	518	607	-1.59	.12	.98	-.3	.94	-.4	.26	37
32	365	607	-.09	.09	1.10	3.2	1.11	1.9	.27	32
3	288	607	.52	.09	1.14	4.2	1.13	2.8	.27	03
23	119	607	2.09	.11	1.12	1.8	1.26	2.5	.28	23
12	516	607	-1.56	.12	.97	-.4	.84	-1.1	.28	12
26	351	607	.02	.09	1.09	3.0	1.05	.9	.29	26
1	476	607	-1.08	.10	.98	-.4	.93	-.6	.30	01
33	362	607	-.06	.09	1.05	1.7	1.02	.3	.32	33
39	538	607	-1.90	.13	.90	-1.1	.64	-2.4	.33	39
2	332	607	.17	.09	1.04	1.5	1.02	.5	.34	02
34	227	607	1.01	.09	1.06	1.6	1.08	1.6	.35	34
29	482	607	-1.15	.11	.93	-1.2	.78	-2.1	.36	29
27	186	607	1.37	.10	1.04	.9	1.10	1.5	.37	27
30	388	607	-.27	.09	.98	-.7	.96	-.5	.37	30
4	240	607	.90	.09	1.03	1.0	1.04	.7	.38	04
21	460	607	-.92	.10	.92	-1.8	.85	-1.7	.38	21
24	171	607	1.52	.10	1.02	.5	1.07	1.1	.38	24
17	396	607	-.34	.09	.96	-1.3	.93	-1.1	.38	17
22	368	607	-.11	.09	.97	-1.0	.93	-1.1	.39	22
25	199	607	1.25	.09	1.00	.1	1.02	.4	.40	25
6	428	607	-.61	.09	.93	-1.9	.80	-2.6	.41	06
13	457	607	-.89	.10	.88	-2.6	.73	-3.1	.43	13
28	425	607	-.59	.09	.88	-3.2	.80	-2.7	.44	28
14	384	607	-.24	.09	.89	-3.5	.80	-3.4	.46	14
5	376	607	-.18	.09	.89	-3.7	.81	-3.3	.47	05
7	421	607	-.55	.09	.86	-4.0	.74	-3.8	.47	07
38	263	607	.71	.09	.86	-4.5	.95	-1.0	.52	38
10	328	607	.20	.09	.85	-5.4	.79	-4.4	.52	10
31	358	607	-.03	.09	.83	-6.1	.75	-4.8	.53	31
11	314	607	.31	.09	.84	-5.8	.79	-4.8	.54	11
8	292	607	.48	.09	.81	-6.6	.77	-5.5	.57	08
9	307	607	.37	.09	.77	-8.4	.72	-6.6	.60	09
MEAN	348.9	607.0	.00	.10	1.00	-.2	.99	-.2		

TABLE 4 - ITEM MEASURE ORDER FOR BBMB1103

Dec 8 17:51 2016

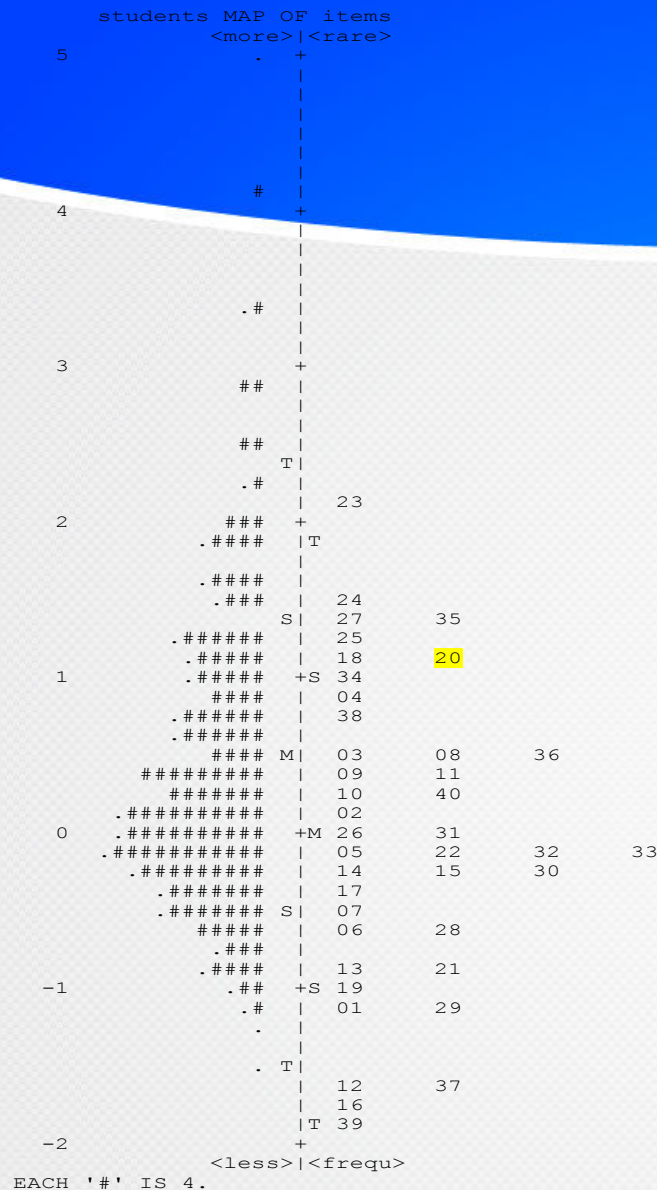
INPUT: 609 students, 40 items MEASURED: 609 students, 40 items, 2 CATS

3.57.3

student: REAL SEP.: 2.19 REL.: .83 ... item: REAL SEP.: 9.51 REL.: .99

item STATISTICS: MEASURE ORDER

ENTRY	RAW			MODEL	INFIT	OUTFIT	PTMEA			
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	item
23	119	607	2.09	.11	1.12	1.8	1.26	2.5	.28	23
24	171	607	1.52	.10	1.02	.5	1.07	1.1	.38	24
27	186	607	1.37	.10	1.04	.9	1.10	1.5	.37	27
35	192	607	1.32	.10	1.27	5.5	1.39	5.8	.16	35
25	199	607	1.25	.09	1.00	.1	1.02	.4	.40	25
18	213	607	1.13	.09	1.24	5.5	1.35	5.8	.18	18
20	218	607	1.09	.09	1.39	8.7	1.53	8.6	.05	20
34	227	607	1.01	.09	1.06	1.6	1.08	1.6	.35	34
4	240	607	.90	.09	1.03	1.0	1.04	.7	.38	04
38	263	607	.71	.09	.86	-4.5	.95	-1.0	.52	38
3	288	607	.52	.09	1.14	4.2	1.13	2.8	.27	03
8	292	607	.48	.09	.81	-6.6	.77	-5.5	.57	08
36	298	607	.44	.09	1.17	5.2	1.20	4.2	.23	36
9	307	607	.37	.09	.77	-8.4	.72	-6.6	.60	09
11	314	607	.31	.09	.84	-5.8	.79	-4.8	.54	11
40	316	607	.30	.09	1.22	7.1	1.31	6.0	.16	40
10	328	607	.20	.09	.85	-5.4	.79	-4.4	.52	10
2	332	607	.17	.09	1.04	1.5	1.02	.5	.34	02
26	351	607	.02	.09	1.09	3.0	1.05	.9	.29	26
31	358	607	-.03	.09	.83	-6.1	.75	-4.8	.53	31
33	362	607	-.06	.09	1.05	1.7	1.02	.3	.32	33
32	365	607	-.09	.09	1.10	3.2	1.11	1.9	.27	32
22	368	607	-.11	.09	.97	-1.0	.93	-1.1	.39	22
5	376	607	-.18	.09	.89	-3.7	.81	-3.3	.47	05
14	384	607	-.24	.09	.89	-3.5	.80	-3.4	.46	14
15	387	607	-.26	.09	1.13	3.9	1.24	3.5	.21	15
30	388	607	-.27	.09	.98	-.7	.96	-.5	.37	30
17	396	607	-.34	.09	.96	-1.3	.93	-1.1	.38	17
7	421	607	-.55	.09	.86	-4.0	.74	-3.8	.47	07
28	425	607	-.59	.09	.88	-3.2	.80	-2.7	.44	28
6	428	607	-.61	.09	.93	-1.9	.80	-2.6	.41	06
13	457	607	-.89	.10	.88	-2.6	.73	-3.1	.43	13
21	460	607	-.92	.10	.92	-1.8	.85	-1.7	.38	21
19	469	607	-1.01	.10	1.03	.7	1.20	1.9	.24	19
1	476	607	-1.08	.10	.98	-.4	.93	-.6	.30	01
29	482	607	-1.15	.11	.93	-1.2	.78	-2.1	.36	29
12	516	607	-1.56	.12	.97	-.4	.84	-1.1	.28	12
37	518	607	-1.59	.12	.98	-.3	.94	-.4	.26	37
16	530	607	-1.77	.13	1.02	.2	1.14	.9	.19	16
39	538	607	-1.90	.13	.90	-1.1	.64	-2.4	.33	39
MEAN	348.9	607.0	.00	.10	1.00	-.2	.99	-.2		
S.D.	107.8	.0	.95	.01	.14	3.8	.21	3.4		



Targetting

Spread of item is wider than spread of students.

Good targetting

No item to measure student at logit 2.

Easy items 12,37,16 and 39.

The target is relatively good. 90% Of the items able to measure 95% of the students this means the items are able to measure about 95% of the students .

Conclusion

- Its fairly a good test, however it needs some improvement on the item –difficult item.

